

A BAYESIAN PARTIAL IDENTIFICATION APPROACH TO INFERRING THE PREVALENCE OF ACCOUNTING MISCONDUCT

P. RICHARD HAHN¹, JARED S. MURRAY², AND IOANNA MANOLOPOULOU³

ABSTRACT. This paper describes the use of flexible Bayesian regression models for estimating a partially identified probability function. Our approach permits efficient sensitivity analysis concerning the posterior impact of priors on the partially identified component of the regression model. The new methodology is illustrated on an important problem where only partially observed data is available— inferring the prevalence of accounting misconduct among publicly traded U.S. businesses.

Keywords: Bayesian inference, nonlinear regression, partial identification, sampling bias, sensitivity analysis, set identification.

1. INTRODUCTION

This paper develops an approach for estimating partially identified parameters in nonlinear regression settings. Our approach is based on a decomposition of the probability function into an identified and a partially identified component [Kadane, 1975]. This representation permits us to employ flexible (nonlinear) models when inferring the identified component; in our applications we utilize Bayesian tree-based priors for the regression functions [Chipman et al., 2010, Hill, 2012]. For the partially identified portion of the model, informative priors are crucial, so checking the sensitivity of posterior inferences to model specification is vital. In our proposed framework, this sensitivity analysis is straightforward, and may be conducted under many different models for the partially identified parameters using only one set of samples from the marginal posterior of the identified parameters.

1. Booth School of Business, University of Chicago.

2. Department of Statistics, Carnegie Mellon University.

3. Department of Statistical Science, University College of London.

The first author thanks the Booth School of Business for supporting this research.

The second author was supported in part by the National Science Foundation under grant numbers SES-11-31897 and SES-1130706. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Our motivating application comes from the corporate accounting literature, where there is substantial interest in determining what fraction of U.S. firms engage in financial misconduct (such as misstated earnings); e.g. Dyck et al. [2013]. Inferring the prevalence of misconduct is complicated by an inherent partial observability—not all cases of misconduct are discovered. Any treatment of this problem will therefore need to analyze how company attributes impact the probability of misconduct being discovered in addition to the probability of the misconduct itself taking place.

As further evidence of the generality of our approach, we also include a reanalysis of a published dataset (from a broken randomized encouragement study of flu vaccine) in the supplementary material.

The remainder of this section collects necessary background material, providing an overview of the concept of partial identification (specifically its treatment from a Bayesian perspective) and describing the empirical data we will analyze. Section 2 lays out our framework more generally and fixes notation. Section 3 describes the results of our data analysis.

1.1. Partial identification. A statistical model $p(y \mid \tau)$ indexed by a parameter $\tau \in \mathcal{T}$ is said to be *identifiable* or *identified* if parameter values correspond uniquely to distinct probability distributions over observables. That is, $p(y \mid \tau) = p(y \mid \tau')$ for all y if and only if $\tau = \tau'$. A model that is not identified is simply referred to as unidentified. The importance of identifiability as a modeling concern has its earliest roots in econometrics, the term first being introduced in Koopmans [1949]. Other seminal references include Haavelmo [1943, 1944] and Koopmans and Reiersol [1950]. Unidentifiability arises naturally in econometric analysis of observational data as a byproduct of imperfect measurement and/or various data censoring mechanisms. The Bayesian perspective on identifiability has been comprehensively reviewed in Aldrich [2002] and more recently in San Martín and González [2010].

The notion of *partial identifiability* or *partial identification* of parameters expands the concept of identification to consider cases of partial learning. A more general definition of identifiability is $p(y \mid \tau) = p(y \mid \tau')$ if and only if $t(\tau) = t(\tau')$ for some non-constant function

t ; here t is an “identifying function” in the terminology of Kadane [1975]. When t is one-to-one, we recover the traditional definition of identifiability, or *point identification*. When the t with the finest preimage satisfying this condition is many-to-one, the model is partially identified — the intuition being that asymptotically we can only isolate the value of $t(\tau)$ consistent with the data, which will correspond to a proper subset of \mathcal{T} with more than one element. For this reason, and in contrast to point identification, it is common to talk of *set identification*. For a more rigorous exposition of the theory of functional identification, refer to Kadane [1975].

Early examples of the partial identification concept include Frisch [1934], Fréchet [1951] and Duncan and Davis [1953]. In recent years, interest in partial identification has accelerated; an excellent recent review article is Tamer [2010] which includes comprehensive citations. See also the book-length treatments by Manski [Manski, 1995, 2003, 2007]. Recent contributions from a Bayesian perspective have focused primarily on asymptotic properties of the posterior distribution over partially identified parameters, notably Gustafson et al. [2005] and Gustafson [2010]. Moon and Schorfheide [2012] examine asymptotic discrepancies between Bayesian credible regions and frequentist confidence sets for set-identified parameters. Florens and Simoni [2011] consider a theoretical framework for studying posteriors of partially identified parameters in nonparametric models. Kline and Tamer [2013] develop large sample approximations of posterior probabilities that particular parameter values lie in the identified set without reference to a prior on the partially identified parameter.

Our approach differs from these recent contributions in three ways. One, it is tailored to a nonlinear regression setting with possibly many predictors and complicated interrelationships; most of the recent literature considers much simpler examples, often without any covariates. Two, our focus is on practical methods for making inferences on parameters of interest with finite samples; most of the recent literature has focused on theoretical and specifically large-sample issues. Three, we introduce an efficient computational scheme for sensitivity analysis, an issue which has received relatively little attention in the literature.

Most previous work focuses on wholly unidentified parameters and typically requires multiple iterations of model fitting; see e.g. McCandless et al. [2007], Molitor et al. [2009] and McCandless et al. [2012] in the context of causal inference/observational data analysis, and Daniels and Hogan [2008] and chapter 15 of Little and Rubin [2002] for extensive reviews in missing data problems.

1.2. Application: inferring the prevalence of accounting misconduct. Since 1982, the United States Securities and Exchange Commission (SEC) has released public notices called Accounting and Auditing Enforcement Releases, or AAERs. AAERs are financial reports “related to enforcement actions concerning civil lawsuits brought by the Commission in federal court and notices and orders concerning the institution and/or settlement of administrative proceedings” [SEC, 2014]. Informally, AAERs comprise a list of publicly traded firms that the SEC has cited for misconduct in one form or another.

For brevity, we adopt the nomenclature “cheating” and “caught”, with the understanding that “cheating” is operationally defined as any accounting anomaly that would lead to an AAER being issued, were it explicitly brought to the SEC’s attention. This interpretation entails that no “caught” firms did not “cheat”, by definition.

Our goal is to provide an estimate of the prevalence of accounting misconduct in the U.S. economy, defined as all actual (caught) and potential (uncaught) AAERs. Predicting which companies are likely to cheat, on the basis of observable firm characteristics, is complicated by the fact that there are potentially many instances of misconduct of which the SEC is unaware. Thus, we do not directly observe which firms cheat, but merely the subset of cheating firms that were *caught* doing so. A naive regression analysis would therefore only speak to the question of which attributes are predictive of getting caught cheating. To complete the analysis, one must incorporate knowledge or conjectures concerning the impact firm attributes have on the likelihood of misconduct being discovered.

2. PARTIALLY OBSERVABLE BINARY RESPONSE VECTORS

Problems with a similar structure to our accounting application appear in the literature under the heading of “partially observed binary data”. Regression models for such data have been studied in many different fields, going by various names. For example, Lancaster and Imbens [1996] considers the case where the observation model is covariate independent under the name “contaminated case-control”, building on Prentice and Pyke [1979]. Poirier [1980] studies such data under the rubric of “partially observed bivariate probit models”, building on the work of Heckman [1976, 1978, 1979]. Our analysis is similar to the approach taken in Wang [2013], which adapts the bivariate probit model of Poirier [1980] for the securities fraud problem.

Whereas these earlier references considered particular parametric models, such as the probit model, and studied identification conditions in that setting, we proceed in the more generic setting of nonlinear regression models, which leads to partially identified parameters. Our approach will be to confront this partial identification with informative priors.

2.1. Reparametrization in partially identified models. As in Dawid [1979], Gelfand and Sahu [1999] and Gustafson et al. [2005], we will work with a reparameterization of τ into an identified component ϕ and an unidentified component (θ, η) . We write the unidentified component as two parameters: θ , which appears in our estimand of interest, and η , which is a nuisance parameter. We will refer to ϕ , $\phi(\mathbf{x})$ or $\phi_{\mathbf{x}}$ (respectively, θ , $\theta(\mathbf{x})$ or $\theta_{\mathbf{x}}$ and η , $\eta(\mathbf{x})$ or $\eta_{\mathbf{x}}$) depending on context. We will use ϕ when the dependence on \mathbf{x} is inessential, we will use $\phi(\mathbf{x})$ to emphasize that ϕ is a function of \mathbf{x} , and we will use $\phi_{\mathbf{x}}$ to refer to point-wise evaluations of $\phi(\mathbf{x})$.

The joint distribution over data and parameters in a partially identified model can be written as

$$\begin{aligned}
(1) \quad \pi(\eta, \theta, \phi, y) &= f(y \mid \eta, \theta, \phi) \pi(\eta, \theta, \phi), \\
&= f(y \mid \phi) \pi(\eta, \theta, \phi), \\
&= f(y \mid \phi) \pi(\phi) \pi(\eta, \theta \mid \phi),
\end{aligned}$$

where the conditional independence implied in moving from the first line to the second line constitutes a definition of partial identification. It follows that the joint posterior distribution of the identified component ϕ and the unidentified component (θ, η) can be written as

$$\begin{aligned}
(2) \quad \pi(\eta, \theta, \phi \mid y) &= \frac{f(y \mid \phi) \pi(\phi) \pi(\eta, \theta \mid \phi)}{f(y)} \\
&= \frac{f(y \mid \phi) \pi(\phi)}{f(y)} \pi(\eta, \theta \mid \phi) \\
&= \pi(\phi \mid y) \pi(\eta, \theta \mid \phi).
\end{aligned}$$

Theorem 5 of Kadane [1975] shows rigorously that the parameter space of any model can be decomposed in this way. Essentially, there are three cases to consider. If the model is fully identified, then (η, θ) is empty, meaning it is a degenerate/constant random variable. When the model has fully unidentified elements, the support of $\pi(\eta, \theta \mid \phi)$ does not depend on ϕ ; the data inform about (η, θ) only via the presumed prior dependence represented in the choice of $\pi(\eta, \theta \mid \phi)$. In the partially identified case, which we focus on here, $\pi(\eta, \theta \mid \phi)$ has support restrictions that do depend on ϕ ; we will denote this ϕ -dependent support by $\Omega(\phi)$.

Our approach will be to directly specify $\pi(\eta, \theta \mid \phi)$ with support $\Omega(\phi)$. The tractability of our approach depends on being able to construct $\pi(\eta, \theta \mid \phi)$ in a way that permits convenient sampling. Our empirical analysis (Section 3.1.1) provides a specific example where this approach is successful (with additional computational details in the appendix).

Although this type of reparameterization is not novel, it has been employed in the literature mainly as a way to study large sample behavior of posterior distributions in partially

identified models; see, for example, Gustafson et al. [2005], Gustafson [2010] and Florens and Simoni [2011]. Most authors suggest using priors on τ to induce the priors on (ϕ, θ) , or simply elect to do indirect inference for θ using only ϕ by computing posteriors over $\Omega(\phi)$. We consider a different approach, working directly in the (ϕ, θ) parameterization and specifying priors as $\pi(\theta | \phi)\pi(\phi)$. This allows us to be fully nonparametric where the data are informative (i.e., on ϕ), and exploit the conditional independence in (2) to conduct computationally efficient sensitivity analysis.

2.2. Partially observed multivariate binary regression models. The general structure of the problem is as follows: The complete data consist of binary vectors $U = (U_1, \dots, U_k)$, of which only certain subsets are simultaneously observable. Interest is in some functional $p(x)$ of the entire joint distribution $p(U_1, \dots, U_k | x)$. Due to the partial observability, $p(x)$ must be reconstructed from an identified function $\phi(x)$ and an unidentified function $\theta(x)$.

Specializing to our corporate accounting analysis, let Z_i indicate “cheating” in firm-year i and let W_i indicate “getting caught” in firm-year i , ($U_1 = W$ and $U_2 = Z$). We assume that with some probability, cheaters get caught, but that there are no firms who get caught when they are not cheating (this is consistent with our operational definition of “cheating”). Additionally, the data are “presence-only” in that we have no confirmation that any given firm is certainly non-cheating.

The parameter of interest is the marginal firm-year probability of cheating,

$$(3) \quad p(x) = \Pr(Z = 1 | x) = \frac{\Pr(W = 1, Z = 1 | x)}{\Pr(W = 1 | Z = 1, x)}$$

from which we may determine the overall prevalence of cheating across all firms as

$$(4) \quad \alpha \equiv \sum_i \Pr(Z_i = 1 | x_i) = \sum_i \frac{\Pr(Z_i = 1, W_i = 1 | x_i)}{\Pr(W_i = 1 | Z_i = 1, x_i)}.$$

Equivalently, for each firm-year we observe $Y_i \equiv Z_i W_i$ instead of (Z_i, W_i) , where Y_i indicating whether a firm received an AAER (cheated and got caught), giving

$$(5) \quad \phi(\mathbf{x}) = \Pr(Z = 1, W = 1 \mid \mathbf{x}) = \Pr(Y = 1 \mid \mathbf{x}); \quad \theta(\mathbf{x}) = \Pr(W = 1 \mid Z = 1, \mathbf{x}).$$

As $\phi(\mathbf{x})$ is simply the (conditional) probability of the observed binary data Y , it is point identified. In our application, we estimate $\phi(\mathbf{x})$ using the BART model described in the Appendix. BART has been shown empirically to be an excellent default nonlinear regression method, with a demonstrated ability to handle many noise variables and strong nonlinearities [Chipman et al., 2010, Hill, 2012].

The partial identification of $p(\mathbf{x})$ arises simply because $0 \leq p(\mathbf{x}) \leq 1$. Given $\phi(\mathbf{x})$, the posterior on $p(\mathbf{x})$ is defined by the prior over $\theta(\mathbf{x})$, truncated to regions satisfying

$$(6) \quad 0 \leq \frac{\phi(\mathbf{x})}{\theta(\mathbf{x})} \leq 1,$$

for all $\mathbf{x} \in \mathcal{X}$. In other words, in our applied setting, $\Omega(\phi(\mathbf{x})) = \{\theta(\mathbf{x}) \mid \phi(\mathbf{x}) \leq \theta(\mathbf{x})\}$.

2.3. A Gaussian process model for the partially identified regression. Furnishing prior information regarding $\theta(\mathbf{x})$ in a predictor-dependent manner strongly motivates the use of simple parametric models. For starters, consider the construction

$$(7) \quad F^{-1}\{\theta(\mathbf{x})\} = h(\mathbf{x})\beta,$$

where F^{-1} is a link function and h denotes some transformation or subset of the covariate vector. A prior over $\theta(\mathbf{x})$ is induced by a prior over β . A chief difficulty with this type of specification is that nonlinear (possibly discontinuous) regression models for $\phi(\mathbf{x})$ impose complex support restrictions on β — indeed, some samples from the posterior for $\phi(\mathbf{x})$ may contradict the model for $\theta(\mathbf{x})$ entirely, meaning that they imply a set of bounds for $\theta(\mathbf{x})$ such that no feasible β exists.

To address this problem, we expand the prior over $\theta(\mathbf{x})$ to acknowledge that (7) is only a guess as to the form of the regression function. Specifically, we *center* our model for $\theta(\mathbf{x})$ at

(7) by assuming that

$$(8) \quad F^{-1}\{\theta(\mathbf{x})\} \mid \phi(\mathbf{x}), \eta(\mathbf{x}) \sim \mathcal{GP}(h(\mathbf{x})\beta, \Sigma_{\mathcal{X}}) \mathbb{1}((\theta(\mathbf{x}), \eta(\mathbf{x})) \in \Omega\{\phi(\mathbf{x})\})$$

for all $\mathbf{x} \in \mathcal{X}$, where $\mathcal{GP}(m, \Sigma_{\mathcal{X}})$ is a Gaussian process with mean m and covariance $\Sigma_{\mathcal{X}}$. In this set-up, β may be fixed or given a prior distribution, in which case we equate $\eta := \beta$.

For the purpose of computing α as in (4), the prior (8) applied to the observed data points reduces to a multivariate normal prior (conditional on β) truncated according to (6). In our empirical analysis, we choose $\Sigma_{\mathcal{X}} = \sigma^2 \mathbf{I}$. Note that this independence is in terms of design points, not in terms of observations: $\theta_{\mathbf{x}_i} = \theta_{\mathbf{x}_j}$ if $\mathbf{x}_i = \mathbf{x}_j$, even if $i \neq j$. That is, draws from the prior (8) are non-smooth functions, as distinct from observation-level errors. Under this simplification, sampling from $(\theta_{\mathbf{x}} \mid \beta, \phi_{\mathbf{x}})$ reduces to drawing samples from independent truncated univariate normal distributions. The more complicated case of a smooth Gaussian process prior (e.g., one with a squared exponential covariance function) would require draws from a multivariate truncated normal distribution; reasonably efficient Hamiltonian Monte Carlo algorithms have been developed for this task [Pakman and Paninski, 2014] but we do not pursue this direction here.

2.4. Simulated example. Before turning to the subject-specific prior we developed for our empirical application, it is instructive to consider how our approach performs in a situation where we know what the right answers ought to be. To this end, we consider a simulated data set based on the competing approach of Wang [2013], which also analyzes the SEC data.

Wang [2013] builds off Poirier [1980], which considers a latent Gaussian utility formulation of the bivariate probit model:

$$(9) \quad \begin{pmatrix} Z^* \\ W^* \end{pmatrix} \sim N(\mu(\mathbf{x}), C_{\rho})$$

where $\mu(\mathbf{x}) = (\gamma_0 + \mathbf{x}^t\gamma, \beta_0 + \mathbf{x}^t\beta)$ and C_ρ is a 2-by-2 correlation matrix with correlation parameter ρ . The observed bivariate binary data is then $Z := \mathbb{1}(Z^* > 0)$ and $W := \mathbb{1}(W^* > 0)$. Specifically, Poirier [1980] establishes that (subject to certain exclusion restrictions) the parameters of the model (γ , β and the correlation parameter ρ) are still identified even when only $Y := ZW$ is observed. Wang [2013] proposes to leverage this result, while deviating from the latent utility formulation. In particular, despite making a “no false positives” assumption (as we do here), Wang [2013] continues to equate $\Pr(Z^* > 0 \mid \mathbf{x})$ with the probability of cheating, which corresponds to the somewhat arbitrary model:

$$\begin{aligned}
& \Pr(\text{cheat, caught}) = \Phi(\gamma_0 + \mathbf{x}^t\gamma, \beta_0 + \mathbf{x}^t\beta, \rho) \\
& \Pr(\text{cheat, not caught}) = \Phi(\gamma_0 + \mathbf{x}^t\gamma) - \Phi(\gamma_0 + \mathbf{x}^t\gamma, \beta_0 + \mathbf{x}^t\beta, \rho) \\
(10) \quad & \Pr(\text{not cheat, caught}) = 0 \\
& \Pr(\text{not cheat, not caught}) = 1 - \Phi(\gamma_0 + \mathbf{x}^t\gamma).
\end{aligned}$$

In other words, Wang [2013] identifies γ from the first equation above, invoking the result of Poirier [1980], and then proceeds to interpret γ as the parameter from a bivariate probit model *without* the no false positives assumption. While there is nothing formally wrong with this model, it would seem to defy justification.

All the same, if (10) is in fact the correct model, it is instructive to observe what our approach gives up to it. Conversely, if (10) is used in a misspecified setting, how do its results compare to ours? To investigate, we simulated $n = 2000$ observations from the following two models. First, we generated data according to (10) by drawing $Y := WZ$ with (W, Z) coming from a bivariate probit model with $\gamma_0 = \beta_0 = -1/2$, $\gamma = (-1, 3/4, 0)$, $\beta = (-3/4, 0, -1/2)$, and $\rho = 1/2$, with \mathbf{x}_1 drawn from a $\text{Uniform}(-\pi/2, \pi/2)$ distribution, and \mathbf{x}_2 and \mathbf{x}_3 drawn from a $\text{Uniform}(-3\pi/2, 3\pi/2)$ distribution (independently). This specification of γ and β satisfies the exclusion restriction of Poirier [1980], in that distinct predictor variables are omitted from each linear equation in the probit mean function.

A Bayesian specification of Wang [2013], with vague conjugate priors for β and γ and a uniform prior on $(-1, 1)$ for ρ , was fit using a Gibbs sampler algorithm with a Metropolis-Hastings update for ρ . Our modular prior approach proceeds by fitting the BART model (with default priors as described in Chipman et al. [2010]) to the observed data (Y_i, \mathbf{x}_i) and constructing the posterior estimate of $\Pr(Z_i | \mathbf{x}_i)$ by dividing posterior samples of $\phi_{\mathbf{X}}$ by draws of $\theta_{\mathbf{X}}$ from (8), with diagonal covariance $\sigma^2 \mathbf{I}$ with $\sigma = 0.1$ and mean function set to match the true $\Pr(W = 1 | Z = 1)$ implied by (10). See the appendix for further computational details regarding model fitting. The results are depicted in Figure 1. As expected, the Wang [2013] model, which achieves point-identification when it is correctly specified, yields much more accurate inference compared to our approach. Meanwhile, even with a correct surveillance model in this case, there persists a modicum of unresolved uncertainty, which reflects that in our model the estimand is only partially identified. Additionally, we see the impact of the BART prior pulling the estimated probabilities towards 1/2 in regions near the edges where there are fewer data points.

Our second demonstration deviates from the linear probit model by specifying $\mu(\mathbf{x}) = (0.5 + \sin(\mathbf{x} + \pi)^t \beta, \sin(\mathbf{x})^t \gamma)$ for the same values of γ and β . The Wang [2013] model is fit with the predictor variables unadjusted. The modular prior approach is fit the same as before, using the correctly specified surveillance model. The results are depicted in Figure 2. As might be expected, under misspecification the Wang [2013] model badly mis-estimates the true $\Pr(Z_i = 1 | \mathbf{x}_i)$. Our approach, with good prior information, still does not achieve point identification, but manages to avoid the gross mis-fit of the Wang [2013] model by successfully recovering the nonlinear identified component from the data.

Naturally, if we had supplied invalid surveillance models, our approach may have been far off the mark in both cases. The point of this demonstration is merely that proceeding in a partially identified fashion is a more conservative course of action than choosing an implausible model on the grounds that — *should it happen to be correct* — it would deliver the desired point identification.

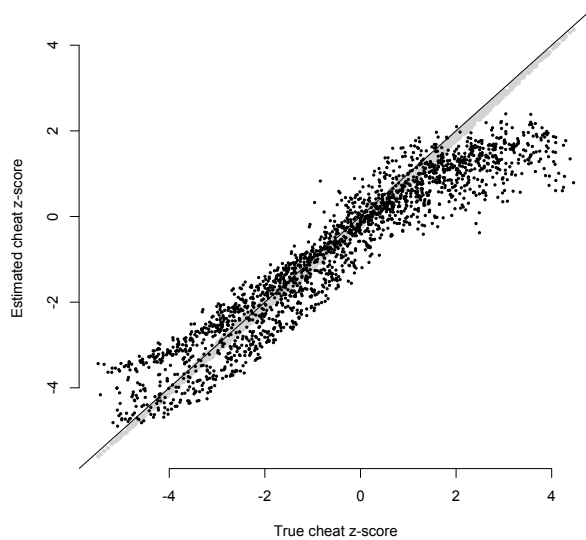


FIGURE 1. Estimated probability of cheating versus the actual probability of cheat on the normal linear predictor scale for each of $n = 2000$ data points. Black solid dots show the modular prior approach, which loosely surrounds the diagonal, demonstrating the unresolved uncertainty in the partial identification approach. The solid gray dots correspond to the Wang [2013] model, which is correctly specified in this example; they hew more tightly to the diagonal.

3. ANALYSIS OF AAER DATA, 2004-2010

Our data are aggregated from three main sources. First, the AAER response variable was obtained from the Center for Financial Reporting and Management (CFRM) at Berkeley’s Haas School of Business. Detailed information about the full data set can be found in Dechow et al. [2011].

Second, additional firm attributes are obtained from the CompuStat North America Annual Fundamentals database via the Wharton Research Data Service (WRDS). These data are then merged with the AAERs using Global Company Key (GVKEY) by year. Specifically, the covariates considered are:

- fiscal year,
- cash,
- net income,
- capital investments,
- SIC industry code,
- qui tam dummy variable.

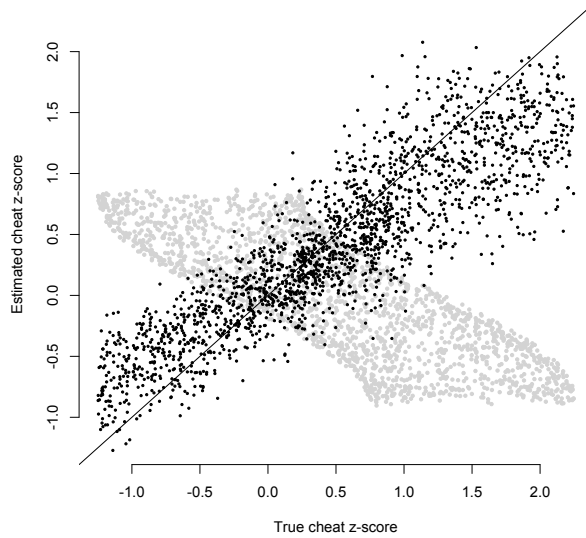


FIGURE 2. Estimated probability of cheating versus the actual probability of cheat on the normal linear predictor scale for each of $n = 2000$ data points. Black solid dots show the modular prior approach, which loosely surrounds the diagonal, demonstrating the unresolved uncertainty in the partial identification approach. The solid gray dots correspond to the Wang [2013] model, which is incorrectly specified in this example; they grossly diverge from the diagonal.

Cash, net income and capital investments are all recorded as a fraction of the firm’s total assets. Standard industrial classifications are given in terms of ten major divisions, denoted A-J by the Occupational Safety and Health Administration. The *qui tam* dummy variable is derived from the SIC codes; it denotes if a firm is in an industry where persons responsible for revealing misconduct are eligible to receive some part of any award resulting from subsequent prosecution. Similar to Dyck et al. [2008] and Jayaraman and Milbourn [2010], our *qui tam* variable is set to one for firms with SIC code 381x, 283x, 37xx, 5122 or 80xx, which includes healthcare providers and pharmaceutical firms, and airplane, missile, and tank manufacturers. It is reasonable to suppose that firms in these industries have a greater likelihood of any misconduct being exposed as a result of incentivized employees.

Finally, a keyword search at the Financial Times web page (www.ft.com) was conducted on each company name and the number of search results was recorded by year. This variable provides a crude measure of media exposure. Although discrepancies between firm names

as recorded in CompuStat and firm names as reported in Financial Times articles lead to measurement error of this variable, it still provides a reasonable proxy for name recognition and cultural visibility. Most firms will never be mentioned in any news article; a few firms are routinely mentioned in the press. To adjust for the fact that web traffic has increased over that period, we normalize the search results count by the total number of hits across all companies in a given year.

We restrict our analysis to U.S. firms who had positive net income for the given year, considering the period between 2004 and 2010, for a total of 6,641 unique firms and a total of $n = 25,889$ total firm-year observations.

3.1. Surveillance model. The unidentified function $\theta(x)$ can be thought of as a “surveillance” probability; it encodes which attributes invite (discourage) SEC scrutiny, making cheaters more (less) likely to be caught. Its effect is to inflate the probability of cheating, which is intuitive since the proportion of caught cheaters $\phi(x)$ represents an obvious lower bound on the proportion of actual cheaters.

Our surveillance model takes the form reported in expression (8) with $\Sigma_{\mathcal{X}} = \sigma^2 \mathbf{I}$ and using a logit link. We scale and shift all variables to reside on the unit interval, taking shifted log transformations of the financial times search hits and net income. We chose to place nonzero coefficients on the (fiscal) year of misconduct, media exposure (as measured by Financial Times search hits), income, cash, and a dummy for qui tam industries. We have specific reasons to expect that these variables are important determinants of the surveillance probability, allowing us to chose informative values for β .

First, note that the frequency of AAERs is substantially higher in earlier years; see Figure 3. AAERs may be filed retroactively, so the opportunity to discover and report misconduct in a given year increases over time. Fitting a curve to the data in Figure 3 suggests a coefficient $\beta_{year} = -2.5$. Observe that this makes the posterior probability of cheating approximately constant across the years examined, which seems plausible.

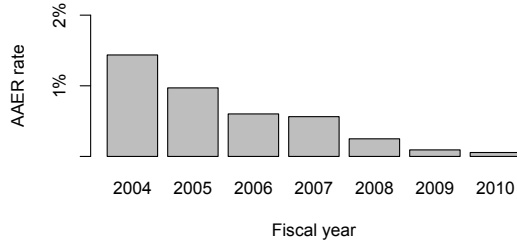


FIGURE 3. AAERs appear more common in more distant years, but this is likely because they may be filed retroactively.

Second, it is reasonable to assume that media attention naturally draws SEC scrutiny [Miller, 2006]. The SEC has a vested interest in catching and making examples of any high-profile cheaters. We set $\beta_{FThits} = 2$, implying approximately a six-fold difference in the probability of misconduct being discovered between a company with no media exposure and the company with the highest media exposure. Similarly, we set $\beta_{quitam} = 1$, implying approximately a two-fold increase in misconduct being discovered for companies in qui tam industries where employees are incentivized to report misconduct.

These observations constitute the subjective information contained in our first observation model (A). To determine the intercept term, consider the following argument: AAERs are quite rare, with an aggregate incidence in our sample of only 0.5%. Potentially this is because very few firms exhibit actionable misconduct, but more likely it is because the SEC has limited resources to identify and pursue violations. Accordingly, one sensible calibration method would be fix the mean probability of discovery across all firms. Fixing this quantity and the other elements of β , we may then solve for the intercept β_0 . In the case of model A, fixing the average discovery rate to 30% gives $\beta_0 = 0$.

After obtaining posterior samples under model A, we observe that cash appears to have a negative impact on the probability of cheating. In contrast, net income appears to have a positive impact on the probability of cheating. We might surmise that these trends are due to unadjusted surveillance probabilities. For example, one could argue that having large amounts of cash on hand (relative to total assets) provides a measure of “wiggle room” that

	Intercept	Fiscal year	FT.com hits	cash	net income	qui tam
β_A	0	-2.5	2	0	0	1
β_B	-0.85	-2.5	2	-1.50	2.5	1

TABLE 1. Regression coefficients defining surveillance models A and B. They differ in their intercept terms and their cash and net income coefficients. The intercepts have been adjusted to obtain an average misconduct discovery rate of 30%

makes certain kinds of misconduct harder to discover. Likewise, firms with high income are more likely to draw SEC attention than firms with smaller income streams. To compare our results under these narratives, we specify a second surveillance model (B), with $\beta_{cash} = -1.5$ and $\beta_{income} = 2.5$. Setting the intercept for model B to match the 30% discovery rate of model A gives $\beta_0 = -0.85$. Surveillance models A and B are shown side-by-side in Table 1.

3.1.1. *Upper bounding cheating.* Finally, we introduce an additional parameter c that allows us to interject prior information concerning an upper bound on the probability of cheating. Recall that the partial identification in this application is driven by the inequality $\Pr(Z = 1 \mid x) = \frac{\phi(x)}{\theta(x)} \leq 1$, which implies $\phi(x) \leq \theta(x)$; the left hand side of this latter inequality is identified by the data. Extremely high probabilities of cheating are implausible, motivating us to consider alternative truncations: $\Pr(Z = 1 \mid x) = \frac{\phi(x)}{\theta(x)} \leq c$, which implies $\phi(x)/c \leq \theta(x)$ for $c \geq \phi(x)$ for all x . Because $\phi(x)$ is identified, the data may contradict any particular value of c by violating $c \geq \phi(x)$ for some x . That is, $\eta := c$ and $\Omega\{\phi(x)\} = \{(c, \theta(x)) \mid \phi(x)/c \leq \theta(x) \leq 1\}$. Note that $\eta := c$ is “global” in that it implies inequalities across all values of x , whereas θ_x can take distinct values at distinct x locations.

We specify $\pi(c \mid \phi(x)) \propto \text{Beta}(10c_0, 10(1 - c_0)) \mathbb{1}\{c \geq \sup_x \phi(x)\}$. The prior mean c_0 captures prior beliefs about the upper bound on $\Pr(Z = 1 \mid x)$, the probability of cheating for any firm. This permits us to specify $\pi(\theta_x \mid c, \phi_x) \propto N(h(x)\beta, \sigma^2) \mathbb{1}\{\theta_x \geq \phi_x/c\}$; joint samples of (c, θ_x) are thereby guaranteed to lie in $\Omega(\phi_x)$.

Our surveillance models allow us to include prior information in the form of subject matter knowledge about the impact of various covariates. We are also able to include additional

subjective prior information about $\sum_i \Pr(W_i = 1 \mid Z_i = 1, \mathbf{x}_i)$ — the overall average probability of a cheating firm getting caught — via the intercept terms, and $\max_i \Pr(Z_i = 1 \mid \mathbf{x}_i)$ — an upper probability on any firm cheating — via the prior on c .

Further computational details are included in the Appendix.

3.2. Results. We conduct a sensitivity analysis by varying the parameters σ and c for both model coefficients β_A and β_B above. Specifically, we consider two settings of each ($\sigma \in \{0.25, 0.5\}$ and $c \in \{0.4, 0.8\}$) for a total of eight candidate models. We study the impact these choices have on both $\Pr(Z = 1 \mid \mathbf{x})$ as a function of individual predictor variables, and also on the overall misconduct prevalence α .

Figure 4 shows the posterior distribution of the adjusted cheating prevalence under the different models. As can be seen in the top panel, increasing σ or c alone has little effect on the overall adjusted cheating prevalence. These two parameters control different aspects of the surveillance uncertainty: a high c implies that any probability of cheating is plausible, whereas a high σ allows large deviations from the specified surveillance model logistic term. Note that with the range of c and σ chosen, neither surveillance model yields overall prevalence of misconduct greater than 15%.

The bottom panel of Figure 4 shows the posterior prevalence for each different SIC code under the two priors, fixing ($\sigma = 0.25, c = 0.4$). Under both priors, SIC category D, representing “electricity, gas, steam and air conditioning supply”, shows much lower cheating prevalence than categories (B,E,H), which correspond to “mining and quarrying”, “water supply, sewerage, waste management and remediation activities”, and “transportation and storage” respectively.

Because BART is nonlinear, summary plots of the impact of individual covariates are challenging to visualize, even if the surveillance model is relatively simple, such as our linear logistic specification. It is instructive, therefore, to examine the implied probability of cheating as one varies individual covariates, for a given firm. That is, how does $\Pr(Z = 1 \mid \mathbf{x})$ change as a function of x_j while holding x_{-j} fixed?

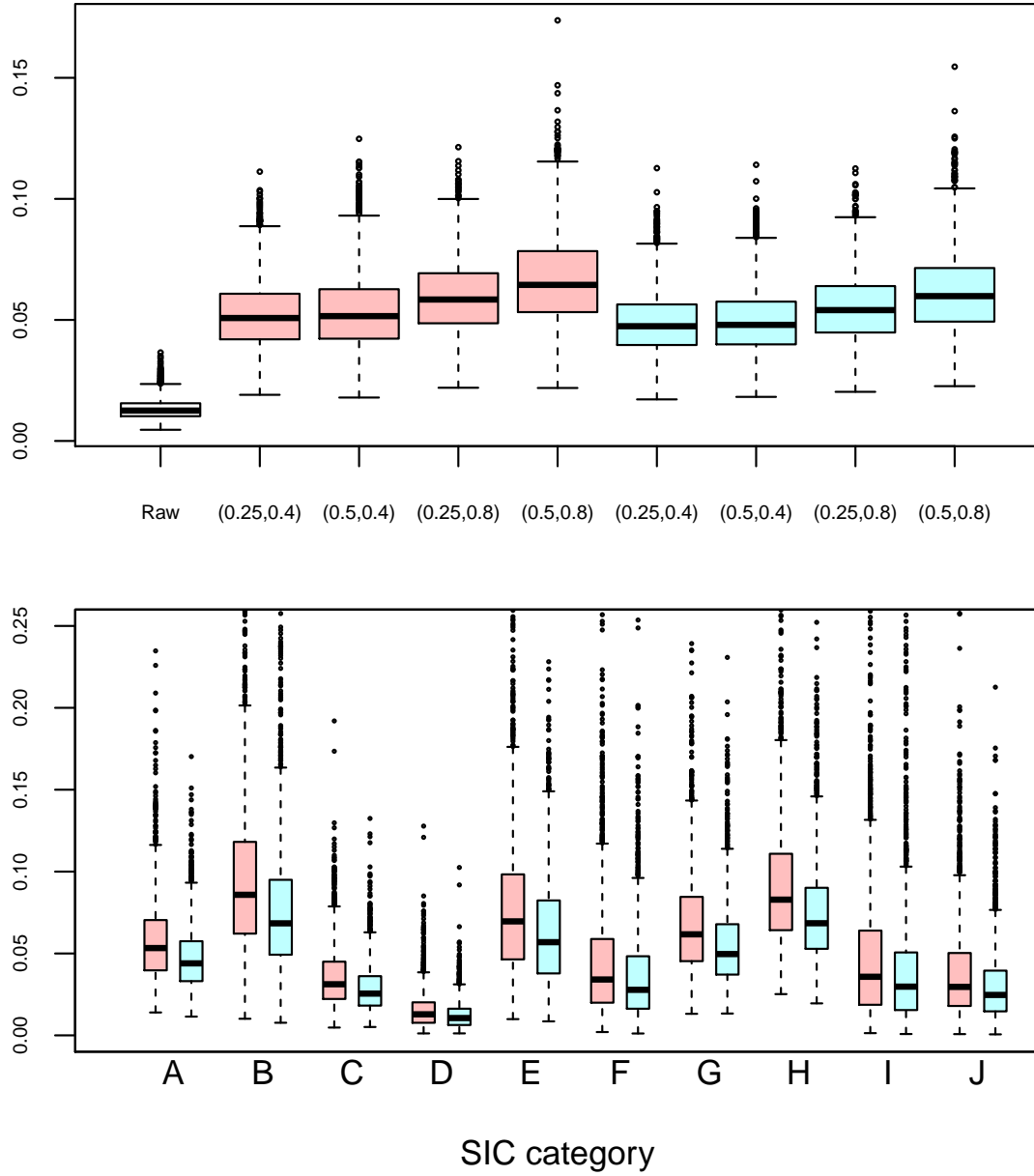


FIGURE 4. Top panel: posterior cheating prevalence, white corresponding to the raw (unadjusted) prevalence, pink to prior A and blue to prior B. The four boxplots within each prior correspond to the following combinations for c and σ (from left to right): $(\sigma = 0.25, c = 0.4)$, $(\sigma = 0.25, c = 0.8)$, $(\sigma = 0.5, c = 0.4)$, $(\sigma = 0.5, c = 0.8)$. Bottom panel: Posterior cheating prevalence in companies within each SIC code, pink corresponding to prior A and blue to prior B for $(\sigma = 0.25, c = 0.4)$.

To demonstrate this approach, we focus on a specific firm, ConAgra Foods of Omaha, Nebraska (simply because it yields illustrative plots). Figure 5 shows $\Pr(Z = 1 \mid \mathbf{x})$, varying media attention, cash, and net income under priors A and B and for various combinations of c and σ . As expected, the surveillance model coefficients on cash and net income reverse the associated slope of the probability of cheating. High values of both σ and c results in posterior credible intervals of up to 40% probability of cheating for some values of net income.

We have reported here only a small number of the possible variations one would presumably want to investigate; we make no claims that models A and B are ideal or even necessarily good or realistic models. Rather, our sensitivity analysis demonstrates a range of possible comparisons that one might undertake when investigating how various assumptions interact with the data via the identified portion of the model.

An important upshot of our analysis is that the surveillance model intercept terms—which govern the average probability of misconduct discovery (getting caught) across firm—and the parameter c —which defines the upper bound $\Pr(Z_i = 1 \mid \mathbf{x}_i) \leq c$ for all i —play dominant roles in determining the inferred overall prevalence of misconduct. For our choices of 30% misconduct discovery probability and $c = 0.4$ or $c = 0.8$, we find that no more than 15% of firms engage in accounting misconduct.

Incidentally, this result is consistent with that of Dyck et al. [2013], who put the prevalence at between 4.74% to 15%, based on a clever natural experiment resulting from the dissolution of the large accounting firm Arthur Andersen and the subsequent re-audit of its clients following the Enron scandal. Unavailability of their exact data, as well as the unavailability of the data of Wang [2013] at the time of writing, means that we cannot compare their precise estimates with those from our model. However, our partial identification analysis suggests that any similar analysis is likely to yield similar conclusions in the matter of overall prevalence. After all, there is only so much information in the available data, with the rest coming from auxiliary assumptions about the surveillance probability, whether those assumptions are explicit, as in our model, implicit, as in the joint likelihood assumed by Wang

[2013], or based on supplementary evidence, as in Dyck et al. [2013]. To the extent that these various approach supply similar assumptions, they will yield similar conclusions. Our approach, by layering such assumptions over the data *ex post*, permits systematic sensitivity analysis rather than one-off comparisons of published studies whose authors are committed to one particular approach.

4. DISCUSSION

Working directly with modular priors in partially identified settings has several advantages. First, it allows identified parameters to be modeled flexibly, permitting the data to be maximally informative, while simultaneously allowing the analyst to specify informative priors for the underidentified components of the model. It may appear that this tactic stands in contrast to the approach of Gustafson [2010] (for example) which advocates working with a scientific model directly in the τ parametrization. However, nothing in our approach precludes the use of such subject-specific information. Rather, we argue that typical prior specifications for τ do not allow separately modulating the prior informativeness on the identified and unidentified components; by working directly in the (ϕ, θ) representation, we achieve precisely this sort differential informativeness. Nonetheless, one should always be mindful of the implied prior on τ . Specifically, analysts can use intuitions regarding τ as a tool for vetting priors over (ϕ, θ) , by checking (via simulation) that they are consistent with available knowledge in the τ representation. In many applications, such as the one studied in this paper, the (ϕ, θ) representation is itself readily interpretable (in this case, the “cheating probability” and the “surveillance probability”, respectively).

Second, when an interpretable parameterization of the modular parameters is available (as it is in our application), the modular prior approach facilitates efficient sensitivity analyses. Sensitivity analysis is good practice generally, and vital when the data are completely uninformative about certain aspects of the model. Being able to conduct such analyses without refitting the entire model can be a tremendous practical advantage, particularly when fitting sophisticated nonlinear regression models to the identified component.

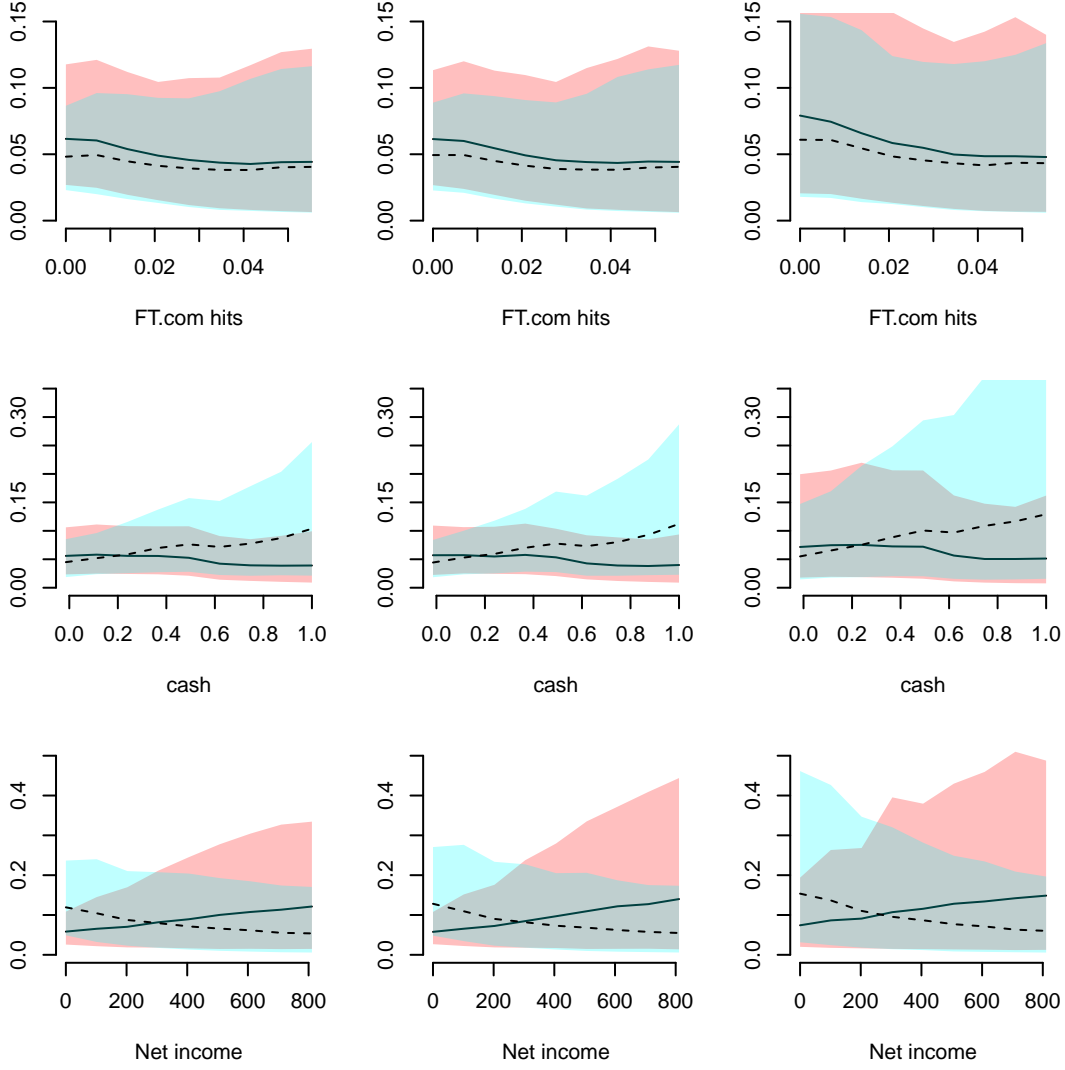


FIGURE 5. $\Pr(Z = 1 | x)$ varying x_j while holding x_{-j} fixed (at the values of ConAgra Foods) under two surveillance models A (pink) and B (blue). Lines depict the median and shaded areas depict 90% credibility intervals. Each row represents a different covariate; the three columns correspond to the following (c, σ) combinations (from left to right): $(c = 0.4, \sigma = 0.25)$, $(c = 0.8, \sigma = 0.25)$, $(c = 0.8, \sigma = 0.5)$.

Finally, while we have not emphasized it in our applications, our use of modular priors allows us to not only conduct sensitivity analyses, but also perform model checking on the unidentified component. By computing “residuals” for our model of $\theta(x)$ based on the support function $\Omega\{\phi(x)\}$ we can learn about areas in the covariate space where our prior

model is contradicted by the data—that is, regions where the model places significant mass outside $\Omega\{\phi(x)\}$. This is a unique opportunity in partially identified models that, to our knowledge, has not been exploited in the existing literature. Performing these model checks in our applications did not raise any red flags, but they would indicate potential problems in situations such as our example in section 2.4.

We expect these methods to be useful in other contexts with partially observed data, including traditional missing data settings (nonignorable unit or item survey nonresponse, dropout, or measurement error) [Rubin, 1976, Little and Rubin, 2002, Daniels and Hogan, 2008], causal inference with observational data or imperfectly randomized study designs [Imbens and Rubin, 1997b, Barnard et al., 2003, Greenland, 2005, McCandless et al., 2007, 2012], and ecological inference [Duncan and Davis, 1953, Phillips Shively, 1969, Achen and Shively, 1995, King, 2013], to name just a few. Our approach allows practitioners to consider much more sophisticated regression modeling for identified parameters while taking a principled approach to the lack of identification, and promotes the routine use of sensitivity analysis and model checking due to the ease with which they are implemented. For an illustration of our approach applied to an imperfectly randomized study of flu vaccine efficacy, see the supplementary material.

REFERENCES

- Securities and Exchange Commission, Accounting and Auditing Enforcement Releases. <http://www.sec.gov/divisions/enforce/friactions.shtml>, 2014.
- C.H Achen and W. Phillips Shively. *Cross-Level Inference*. University Of Chicago Press, first edition, 5 1995.
- J. Aldrich. How likelihood and identification went Bayesian. *International Statistical Review*, 70(1):79–98, 2002.
- J.D. Angrist, G.W. Imbens, and D.B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.
- J. Barnard, C.E. Frangakis, J.L. Hill, and D.B. Rubin. Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in New York City. *Journal of the American Statistical Association*, 98(462):299–323, 2003.
- J.C.C. Chan and J.L. Tobias. Priors and posterior computation in linear endogenous variable models with imperfect instruments. *Journal of Applied Econometrics*, 2014. ISSN 1099-1255.

- H.A. Chipman, E.I. George, and R.E. McCulloch. Bayesian CART model search. *Journal of the American Statistical Association*, 93(443):935–948, 1998.
- H.A. Chipman, E.I. George, and R.E. McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.
- D.R. Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 215–242, 1958.
- M.J. Daniels and J.W. Hogan. *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*. CRC Press, 2008.
- A.P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B*, 41(1):1–31, 1979.
- P.M. Dechow, W. Ge, C.R. Larson, and R.G. Sloan. Predicting material accounting misstatements. *Contemporary accounting research*, 28(1):17–82, 2011.
- O.D. Duncan and B. Davis. An alternative to ecological correlation. *American Sociological Review*, 18(6):665–666, 1953.
- A. Dyck, N. Volchkova, and L. Zingales. The corporate governance role of the media: Evidence from Russia. *The Journal of Finance*, 63(3):1093–1135, 2008.
- A. Dyck, A. Morse, and L. Zingales. How pervasive is corporate fraud? Technical Report 2222608, Rotman School of Management, 2013.
- J.-P. Florens and A. Simoni. Bayesian identification and partial identification. *Unpublished manuscript. Toulouse School of Economics, Toulouse, France*, 2011.
- M. Fréchet. Sur les tableaux de corrélation dont les marges sont données. *Ann. Univ. Lyon Sect. A*, 9:53–77, 1951.
- R. Frisch. *Statistical Confluence Analysis by Means of Complete Regression Systems*. Universitetets Økonomiske Institut, 1934.
- A.E. Gelfand and S.K. Sahu. Identifiability, improper priors and Gibbs sampling for generalized linear models. *Journal of the American Statistical Association*, 94(445):247–253, 1999.
- S. Greenland. Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(2):267–306, 2005.
- P. Gustafson. Bayesian inference for partially identified models. *The International Journal of Biostatistics*, 6(2), 2010.
- P. Gustafson, A.E. Gelfand, S.K. Sahu, W.O. Johnson, T.E. Hanson, L. Joseph, and J. Lee. On model expansion, model contraction, identifiability and prior information: Two illustrative scenarios involving mismeasured variables [with comments and rejoinder]. *Statistical Science*, pages 111–140, 2005.
- T. Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica*, 11(1):pp. 1–12, 1943.
- T. Haavelmo. The probability approach in econometrics. *Supplement to Econometrica*, 12: iii–115, July 1944.
- J.J. Heckman. Simultaneous equation models with continuous and discrete endogenous variables and structural shifts. In S.M. Goldfeld and R.E. Quandt, editors, *Studies in nonlinear estimation*, pages 235–272. Balinger, 1976.
- J.J. Heckman. Dummy endogenous variables in a simultaneous equation system. *Econometrica*, 46:931–959, 1978.
- J.J. Heckman. Sample bias as a specification error. *Econometrica*, 47:153–161, 1979.

- J.L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):212–240, 2012.
- K. Hirano, G.W. Imbens, D.B. Rubin, and X.H. Zhou. Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, 1(1):69–88, March 2000.
- G. W. Imbens and D. B. Rubin. Estimating Outcome Distributions for Compliers in Instrumental Variables Models. *The Review of Economic Studies*, 64(4):555–574, October 1997a. ISSN 0034-6527. doi: 10.2307/2971731. URL <http://restud.oxfordjournals.org/lookup/doi/10.2307/2971731>.
- G.W. Imbens and D.B. Rubin. Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, pages 305–327, 1997b.
- S. Jayaraman and T. Milbourn. Whistle Blowing and CEO Compensation: The Qui Tam Statute. In *AFA 2011 Denver Meetings Paper*, 2010.
- J. Kadane. The role of identification in Bayesian theory. In S.E. Fienberg and A. Zellner, editors, *Studies in Bayesian econometrics and statistics*, chapter 5.2, pages 175–191. North-Holland, 1975.
- G. King. *A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data*. Princeton University Press, 2013.
- B. Kline and E. Tamer. Default Bayesian inference in a class of partially identified models. *manuscript, Northwestern University*, 2013.
- T.C. Koopmans. Identification problems in economic model construction. *Econometrica*, pages 125–144, 1949.
- T.C. Koopmans and O. Reiersol. The identification of structural characteristics. *Annals of Mathematical Statistics*, 21(2):165–181, 1950.
- T. Lancaster and G. Imbens. Case-control studies with contaminated controls. *Journal of Econometrics*, 71:145–160, 1996.
- R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data*. Wiley-Interscience, 2 edition, 9 2002. ISBN 9780471183860.
- C.F. Manski. *Identification problems in the social sciences*. Harvard University Press, 1995.
- C.F. Manski. *Partial identification of probability distributions*. Springer, 2003.
- C.F. Manski. *Identification for prediction and decision*. Harvard University Press, 2007.
- L.C. McCandless, P. Gustafson, and A. Levy. Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Statistics in medicine*, 26(11):2331–2347, 2007.
- L.C. McCandless, P. Gustafson, A.R. Levy, and S. Richardson. Hierarchical priors for bias parameters in Bayesian sensitivity analysis for unmeasured confounding. *Statistics in medicine*, 31(4):383–396, 2012.
- C.J. McDonald, S.L. Hui, and W.M. Tierney. Effects of computer reminders for influenza vaccination on morbidity during influenza epidemics. *MD computing: computers in medical practice*, 9(5):304–312, 1991.
- Gregory S Miller. The press as a watchdog for accounting fraud. *Journal of Accounting Research*, 44(5):1001–1033, 2006.
- N.-T. Molitor, N. Best, C. Jackson, and S. Richardson. Using Bayesian graphical models to model biases in observational studies and to combine multiple sources of data: application to low birth weight and water disinfection by-products. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(3):615–637, 2009.

- H.R. Moon and F. Schorfheide. Bayesian and frequentist inference in partially identified models. *Econometrica*, 80(2):755–782, 2012.
- Ari Pakman and Liam Paninski. Exact hamiltonian monte carlo for truncated multivariate gaussians. *Journal of Computational and Graphical Statistics*, 23(2):518–542, 2014.
- W. Phillips Shively. “Ecological” inference: The use of aggregate data to study individuals. *The American Political Science Review*, pages 1183–1196, 1969.
- D.J. Poirier. Partial observability in bivariate probit models. *Journal of Econometrics*, 12(2):209–217, 1980.
- R.L. Prentice and R. Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411, 1979.
- T.S. Richardson and J.M. Robins. Analysis of the binary instrumental variable model. In Rina Dechter, Hector Geffner, and Joseph Y. Halpern, editors, *Heuristics, Probability and Causality*, chapter 25. College Publications, 2010.
- T.S. Richardson, R.J. Evans, and J.M. Robins. Transparent parameterizations of models for potential outcomes. In J.M. Bernardo, M.J. Bayarri, James O. Berger, A. P. Dawid, David Heckerman, A.F.M Smith, and Mike West, editors, *Bayesian Statistics*, volume 9, pages 569–610. Oxford University Press, 2011.
- D.B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.
- E. San Martín and J. González. Bayesian identifiability: contributions to an inconclusive debate. *Chilean Journal of Statistics*, 1(2):69–91, 2010.
- E. Tamer. Partial identification in econometrics. *Annual Review of Economics*, 2:167–95, 2010.
- T.Y. Wang. Corporate securities fraud: insights from a new empirical framework. *Journal of Law, Economics and Organization*, 29(3):535–568, 2013.
- J.L. Zhang and D.B. Rubin. Estimation of causal effects via principal stratification when some outcomes are truncated by “death”. *Journal of Educational and Behavioral Statistics*, 28(4):353–368, 2003.

APPENDIX A. BAYESIAN ADDITIVE REGRESSION TREES

In our application, the nonlinear function $\phi(x)$ was modeled using the Bayesian Additive Regression Trees (BART) approach of Chipman et al. [2010]. Though nothing in our method is specific to this choice, the BART model has several properties that make it a sensible choice. BART is more flexible than classical parametric regression models, such as linear logistic or probit regression [Cox, 1958], but, unlike alternative nonparametric Bayesian regression models, BART is able to detect interactions and discontinuities and is invariant to monotone transformations of the covariates. For clarity, in this appendix we keep to the original notation used in Chipman et al. [2010]; several of the BART prior parameter names are used in the main text where they refer to different entities.

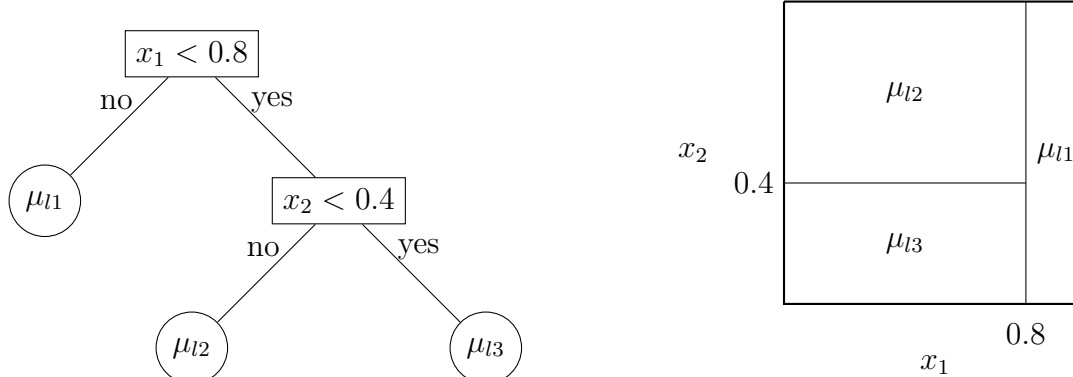


FIGURE 6. (Left) An example binary tree, with internal nodes labelled by their splitting rules and terminal nodes labelled with the corresponding parameters μ_{lb} (Right) The corresponding partition of the sample space and the step function.

The BART prior represents an unknown function $f(x)$ as a sum of many piecewise constant binary regression trees. Each tree T_l , $1 \leq l \leq L$, consists of a set of internal decision nodes which define a partition of the covariate space (say $\mathcal{A}_1, \dots, \mathcal{A}_{B(l)}$), as well as a set of terminal nodes or leaves — one corresponding to each element of the partition. Each subset of the partition \mathcal{A}_b is associated with a parameter value μ_{lb} , defining a piecewise constant function: $g_l(x) = \mu_{lb}$ if $x \in \mathcal{A}_b$. This regression tree function representation is depicted in Figure 6.

Individual regression trees are then additively combined into a single regression function: $f(x) = \sum_{l=1}^L g_l(x)$. The representation of $f(x)$ through the sum of a set of regression trees is generally non-unique; in our applications, this redundancy is unproblematic.

Each of the functions g_l are constrained by their prior to be “weak learners”; that is, the prior strongly favors small trees and leaf parameters that are near zero. Each tree independently follows the prior described by Chipman et al. [1998], where the probability that a node at depth d splits (is not terminal) is given by $\alpha(1+d)^{-\beta}$, $\alpha \in (0, 1)$, $\beta \in [0, \infty)$.

A variable to split on, and a cut-point to split at, are then selected uniformly at random from the available splitting rules. Large, deep trees are given extremely low prior probability by taking $\alpha = 0.95$ and $\beta = 2$ as in Chipman et al. [2010]. The leaf parameters are assigned independent priors $\mu_{lb} \sim N(0, \sigma_\mu^2)$ where $\sigma_\mu = 3/(k\sqrt{L})$. The default value, $k = 2$, shrinks $g_l(x)$ strongly toward zero. The induced prior for $f(x)$ is centered at zero and

puts approximately 95% of the prior mass within ± 3 pointwise. Larger values of k imply increasing degrees of shrinkage.

Finally, the probability function of interest is modeled as $\Phi(f(\mathbf{x}) + c)$, where c is an offset parameter and $\Phi(\cdot)$ is the standard normal cumulative distribution function. Complete details of the BART prior and its implementation are given by Chipman et al. [2010].

A.1. Computation. Posterior inferences are obtained by sampling at a fixed grid of design points $\mathbf{x}_1^*, \dots, \mathbf{x}_J^*$. To reduce notational clutter, we suppress dependence on \mathbf{x}_j^* ; for example, θ should be read as referring to a specific $\theta(\mathbf{x}_j^*)$.

Operationally, posterior samples are obtained according to the following recipe.

- (1) Fit the BART model to the observed pairs (Y_i, \mathbf{x}_i) , for $i = 1, \dots, n$. Convenient R implementations facilitate this step readily, for example `BayesTree` or `dbarts`. This gives a collection of posterior samples ϕ^1, \dots, ϕ^k .
- (2) For each posterior sample ϕ^k , draw θ^k by first drawing η^k from $\pi(\eta^k \mid \phi^k) \mathbb{1}\{\eta^k \in \Omega[\phi^k]\}$ and then drawing θ^k from $\pi(\theta^k \mid \phi^k, \eta^k) \mathbb{1}\{\theta^k, \eta^k \in \Omega[\phi^k]\}$.

See Chan and Tobias [2014] for a similar computational approach. Because sampling from the posterior of ϕ operates independently from sampling the partially identified parameter θ — as shown in (2) in the main text — sensitivity analysis can be conducted without ever needing to refit the model, simply by repeating step 2 for various choices of $\pi(\eta, \theta \mid \phi)$.

In our empirical application, we use the Gaussian process prior (8) with diagonal covariance matrix, $\eta := c$ is given a Beta prior as described in Section 3.1.1, and we use a standard normal probit link for F^{-1} . For these choices, step 2 above becomes:

- (i) Draw c from its truncated Beta distribution with lower truncation point given by $\max\{\phi(\mathbf{x}_1^*), \dots, \phi(\mathbf{x}_J^*)\}$.
- (ii) Draw $F^{-1}\{\theta_{\mathbf{x}}\}$ from independent truncated normal distributions at each design point, with lower bound $F^{-1}\{\phi_{\mathbf{x}}/c\}$.

APPENDIX B. SUPPLEMENTARY APPLICATION: RANDOMIZED ENCOURAGEMENT STUDY OF FLU VACCINE

Our data come from a randomized encouragement study of the effectiveness of an influenza vaccine conducted at an academic primary care practice affiliated with a large teaching hospital in an urban setting [McDonald et al., 1991]. Physicians were randomly assigned to treatment or control groups, with treated physicians receiving a computerized reminder about the vaccine when a patient with a scheduled visit also met the U.S. Public Health Service Criteria for receiving a flu vaccine. Patients were eligible for the vaccine if they were 65 or older, or if they had certain severe chronic diseases including chronic obstructive pulmonary disease (COPD). The outcome of interest is subsequent hospitalization for flu-related illness, and the covariates include the patient’s age and COPD status. Hirano et al. [2000] conduct a Bayesian analysis of this problem, placing a default joint prior over all the identified and unidentified parameters. In this application we perform an alternative selectively informative analysis with modular priors. We will use the terms “reminder”, “instrument” and “encouragement” interchangeably here, preferring “reminder” in the context of the specific applied problem.

B.1. Causal Inference in a Binary Instrumental Variables Model. Let Z_i, T_i , and Y_i be the observed values of the treatment assignment (or encouragement, or instrument), treatment received, and outcome for observation i , respectively. Let X_i be a vector of baseline covariates. In the flu data, $Z_i = 1$ if encouragement (the reminder) was administered, $T_i = 1$ if the vaccine was administered, and $Y_i = 1$ if the patient was *not* hospitalized. As in Angrist et al. [1996] and Hirano et al. [2000] define the *potential outcomes* $T_i(z)$ and $Y_i(t, z)$ as the treatment assignment that would have been observed had we fixed $Z_i = z$ and the outcome we would have observed fixing $T_i = t, Z_i = z$, respectively. We will suppress indexing by i for population quantities. We make the stable unit treatment value assumption (SUTVA) [Angrist et al., 1996], i.e. that there is no interference between units and no

$T(1)$	$T(0)$	S (Compliance Type)	
0	0	Never Taker	(n)
1	1	Always Taker	(a)
1	0	Complier	(c)
0	1	Defier	(d)

TABLE 2. Compliance types, or principal strata, defined by response patterns across levels of Z .

alternative versions of treatment.¹ Mathematically, SUTVA implies that $T_i(\mathbf{z}) = T_i(z_i)$ and $Y_i(\mathbf{t}, \mathbf{z}) = Y_i(t_i, z_i)$, where \mathbf{z}, \mathbf{t} are vectors of instrument/encouragement assignment and treatments received (respectively) for the entire sample.

In this application the complete vector of binary response variables is

$$(11) \quad U_i = (T_i(0), T_i(1), Y_i(0, 0), Y_i(0, 1), Y_i(1, 0), Y_i(1, 1))',$$

the collection of treatments and outcomes under all possible experimental conditions. However, we observe only $T_i \equiv T_i(Z_i)$ and $Y_i \equiv Y_i(T_i(Z_i), Z_i)$, where Z_i is randomly assigned at the study outset. We can define compliance types or principal strata S based on the possible values for $(T(1), T(0))$ – see Table 2. We will assume that the encouragement Z is ignorable or randomly assigned (as it is in the flu data). We also assume that there are no defiers (“monotonicity of compliance”), i.e., $\Pr(S = d \mid \mathbf{x}) = 0$. In the current setting a defier would be a patient who would *only* receive the vaccine if the reminder was *not* administered, which is quite unlikely.

Let $\pi_s(x) = \Pr(S = s \mid \mathbf{x})$ and $\gamma_s^{tz}(x) = \Pr(Y(t, z) = 1 \mid S = s, \mathbf{x})$ denote the principal strata and potential outcome probability models. Under the above assumptions the potential outcome and principal strata distributions are linked to the joint distribution of observables

¹Hirano et al. [2000] point out that in an infectious disease application with individuals as the unit of randomization this assumption is unlikely to hold exactly, but constitutes a reasonable first order approximation.

by the following equations.

$$(12) \quad \Pr(T = 1 \mid Z = 0, \mathbf{x}) = \pi_a(\mathbf{x}),$$

$$(13) \quad \Pr(T = 1 \mid Z = 1, \mathbf{x}) = \pi_c(\mathbf{x}) + \pi_a(\mathbf{x}),$$

$$(14) \quad \Pr(Y = 1 \mid T = 0, Z = 0, \mathbf{x}) = [\pi_c(\mathbf{x})\gamma_c^{00}(\mathbf{x}) + \pi_n(\mathbf{x})\gamma_n^{00}(\mathbf{x})]/[\pi_c(\mathbf{x}) + \pi_n(\mathbf{x})],$$

$$(15) \quad \Pr(Y = 1 \mid T = 1, Z = 1, \mathbf{x}) = [\pi_c(\mathbf{x})\gamma_c^{11}(\mathbf{x}) + \pi_a(\mathbf{x})\gamma_a^{11}(\mathbf{x})]/[\pi_c(\mathbf{x}) + \pi_a(\mathbf{x})],$$

$$(16) \quad \Pr(Y = 1 \mid T = 0, Z = 1, \mathbf{x}) = \gamma_n^{01}(\mathbf{x}),$$

$$(17) \quad \Pr(Y = 1 \mid T = 1, Z = 0, \mathbf{x}) = \gamma_a^{10}(\mathbf{x}).$$

Richardson and Robins [2010] derive this mixture relationship between principal strata/potential outcome distributions and the observable joint in detail under a range of assumptions about the latent variables (see also Imbens and Rubin [1997a] and Richardson et al. [2011]).

The probabilities on the left hand side completely determine the joint distribution of observables (Y and T given Z) and hence constitute a basis for ϕ . Clearly π , γ_n^{01} and γ_a^{10} are one-to-one maps from ϕ , and are point identified as well. However, γ_c^{00} , γ_n^{00} , γ_c^{11} , and γ_a^{11} are not. For example, at any point in covariate space $\gamma_c^{11}(x)$ and $\gamma_a^{11}(x)$ must lie on the line segment given by the intersection of the line in equation (15) and $[0, 1] \times [0, 1]$, but the data are wholly uninformative about their position on that line segment. In particular, the range of $\gamma_a^{11}(x)$ is

$$(18) \quad \left(\min \left(0, \frac{\pi_a(x) + \pi_c(x)}{\pi_a(x)} p_{1|11}(x) - \frac{\pi_c(x)}{\pi_a(x)} \right), \max \left(1, \frac{\pi_a(x) + \pi_c(x)}{\pi_a(x)} p_{1|11}(x) \right) \right)$$

where $p_{1|11}(x) = \Pr(Y = 1 \mid T = 1, Z = 1, \mathbf{x})$. The situation is exactly the same for $\gamma_c^{00}(x)$ and $\gamma_n^{00}(x)$, with the line given by (14) and bounds similar to (18) (large sample bounds on causal effects within principal strata were introduced by Zhang and Rubin [2003]; see Richardson et al. [2011] for bounds under monotonicity and other assumptions, as well as further references).

Point identification for the remaining parameters is often obtained through further assumptions on the distribution of the potential outcomes. The *stochastic exclusion restriction* for compliance type s posits that there is no direct effect of Z on Y ; that is, $\gamma_s^{i0} = \gamma_s^{i1}$ for $i = 0, 1$ [Hirano et al., 2000]. From (12)-(17) it is clear that this achieves point identification for the entire distribution of the potential outcomes.

Hirano et al. [2000] anticipated failures of the exclusion restrictions, noting that a reminder about flu shots might also prompt physicians to counsel patients about other preventative measures to avoid contracting the flu. They propose relaxing the exclusion restrictions and resolving the resulting nonidentifiability through prior probability modeling. In this sense our approach and theirs are philosophically similar, but our framework affords us significant flexibility in choosing the model for ϕ while transparently embedding prior information about θ , which is difficult to do in their modeling setup. In our subsequent analysis we will avoid imposing any exclusion restrictions.

B.2. Model Specification. Our model for observables decomposes into a treatment model and an outcome given treatment model, i.e., $P(Y, T \mid Z, \mathbf{x}) = P(T \mid Z, \mathbf{x})P(Y \mid T, Z, \mathbf{x})$. We assign each of these a probit BART prior. For the treatment model, randomization of treatment assignment and monotonicity of compliance imply that $\Pr(T = 1 \mid Z = 1, \mathbf{x}) \geq \Pr(T = 1 \mid Z = 0, \mathbf{x})$ for all \mathbf{x} (see equations (12) and (13)). We enforce this constraint by rejecting samples from BART where this condition fails at any point in covariate space. More elaborate approaches to enforcing monotonicity are possible but beyond the scope of the present paper.

B.2.1. Partially identified parameters. Recall that we have two pairs of unidentified parameters, $(\gamma_c^{00}, \gamma_n^{00})$ and $(\gamma_c^{11}, \gamma_a^{11})$. Fixing one element of each pair would solve the identification problem, so we can take $\theta = (\gamma_n^{00}, \gamma_a^{11})$ without loss of generality. This is in contrast to Hirano et al. [2000], who specify a strongly parametric probability model for all four of the unidentified parameters.

We consider three priors on θ . All three are of the form

$$(19) \quad (\text{logit}\{\gamma_n^{00}(x)\}, \text{logit}\{\gamma_a^{11}(x)\})' = \mu\{\phi(x)\} + h(x)\beta + \epsilon(x).$$

with $\mu\{\phi(x)\} = (\text{logit}\{\gamma_n^{01}(x)\}, \text{logit}\{\gamma_a^{10}(x)\})'$, $h(x) = 1$ and $\beta = (b_a, -b_n)$. This centers our prior at a proportional odds model, where the reminder increases the odds of a positive outcome by a factor of $\exp(b_a)$ or $\exp(b_n)$ in always or never takers (respectively) at any covariate value. A natural way to complete this prior is to take $b_s \sim N(0, \sigma_s^2)$ for $s = a, n$ which centers the prior at the exclusion restrictions and expresses prior beliefs about plausible effect sizes through σ_s^2 . For illustration we take $\sigma_s = 0.1$ or 0.25 for the current application. Hirano et al. [2000] estimate principal strata-independent total effects of COPD around -0.35 on the log odds scale, so the more diffuse prior covers a plausible range of possible effect sizes for the direct effect of a reminder within non-compliers. We would expect any direct beneficial effect of the reminder on the subsequent hospitalization to be smaller than the detrimental effect of a serious chronic disease like COPD purely on substantive grounds.

In the flu application we also have some prior information about likely effect directions, which we can also incorporate in the prior. Any direct effect from the reminder is most likely to be positive, since the reminder probably encourages doctors and patients to take other preventative measures beyond the vaccine. Additionally, Hirano et al. [2000] speculate that the exclusion restriction for never takers is more plausible than for always takers since perceived risk is a key driver of compliance behavior, with low-risk patients more likely to be never takers. A more nuanced view is that the direct effect should be *larger* for always takers than never takers — even though never takers probably have lower risk on average, we should not rule out the possibility of high-risk never takers who refuse the vaccine due to cost, safety or other concerns.

A prior for θ that reflects these beliefs is $(b_a, b_n) \sim N(b_0, \Sigma_b)\mathbb{1}(b_a > 0, b_n > 0)$. For the analysis in the sequel we will take $b_0 = (0.05, 0.0)'$, $\Sigma_b = \text{diag}(0.13^2, 0.05^2)$. In additional simulations we find that results are insensitive to plausible values of $\text{Cov}(b_a, b_n)$. These

particular numerical values were chosen. The prior mode is b_0 , the prior mean is about $(0.125, 0.04)$, and the marginal 90th percentile is about 0.25 for b_a and 0.08 for b_n . The prior scale for b_a is calibrated so that (with high probability) b_a is smaller in magnitude than the effect of COPD on the hospitalization as reported by Hirano et al. [2000] (the posterior mean was generally around -0.35, SD 0.15). We will call this the “informed” prior. For all three priors we set $v = 0.025$.

B.3. Results. We fit the model described in the previous section to the flu vaccine data, using age and COPD as covariates. Figure 7 shows the probability of belonging to each principal stratum as a function of age and COPD along with marginal 90% credible intervals. We find a fairly sharp break in compliance behavior at age 65 (the probability of being a never-taker drops, and seems to be shifted largely to always takers), which appears to hold regardless of COPD status. Possible explanations include patients or providers giving more weight to age-based vaccination guidelines, or better insurance coverage and increased access to services in the older cohort.

Figure 8 shows the outcome distributions by age, COPD, and the treatment assigned and received. There is clear evidence of an age/COPD interaction; this is due at least in part to the study design, where all participants under 65 have at least one chronic disease (possibly COPD). This also means that the 65+ population is on average less likely to be hospitalized than the younger cohort, explaining the seemingly paradoxical decrease in the estimated probability of hospitalization at age 65 in the “No COPD” group.

B.3.1. Causal effects. We are primarily interested in (partially identified) sample causal effects. As in Hirano et al. [2000] we will consider estimating the intention-to-treat (ITT) effect in each principal stratum:

$$(20) \quad ITT_s = \frac{1}{n_s} \sum_{i: S_i=s} (Y_i(T_i(1), 1) - Y_i(T_i(0), 0)).$$

where n_s is the number of observations with compliance type S . Given a sample from the marginal distribution for ϕ , and a sample from θ , we can impute each observation’s missing

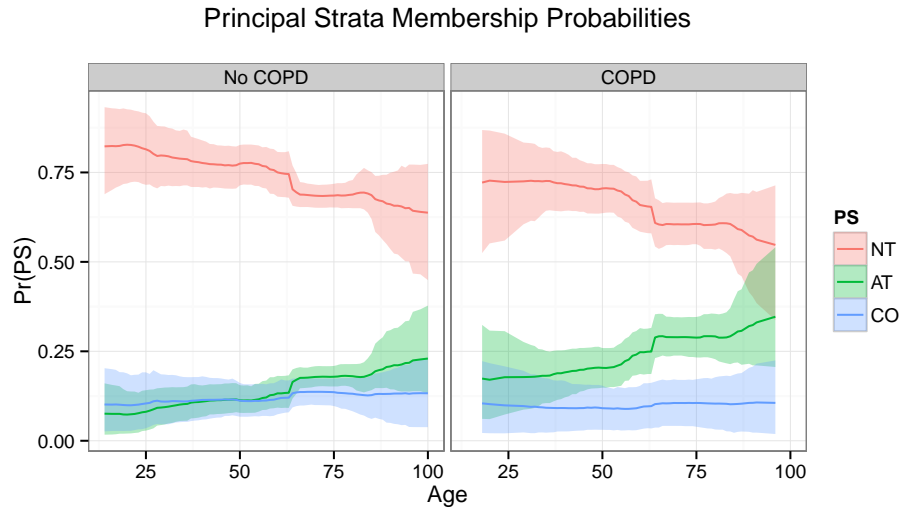


FIGURE 7. Principal strata membership probabilities by age and COPD, including marginal 90% credible intervals

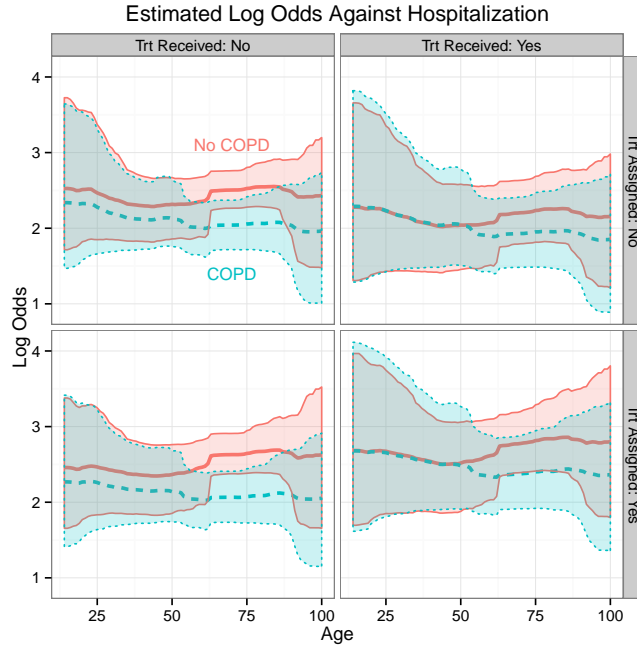


FIGURE 8. Log odds of a positive outcome (no hospitalization) by age, COPD, and treatment assignment/receipt

principal strata membership and the unobserved potential outcome. With the completed data it is straightforward to compute the sample ITT effects. Computational details are provided in Appendix C

We compare results under our three priors to three models from Hirano et al. [2000]: the model with no exclusion restrictions (“none”), both exclusion restrictions (“both”), or the never-taker exclusion restriction (“n”). The “Hirano (none)” model has very wide standard deviations, which are driven by their use of diffuse, “off-the-shelf” priors. Their marginal prior for the direct effect of the reminder is approximately normal with mean zero and standard deviation 1.5, which puts significant prior probability on implausible values such as large and/or negative direct effects. Such vague priors are inappropriate for partially identified parameters.

In contrast, the BART models all give similar results. The “BART (0.1)” and “BART (0.25)” models differ mainly in the width of the posterior standard errors, and then only slightly. The “BART (0.25)” model assumes that with high probability the magnitude of any direct effect on the log odds is within ± 0.5 – compare this to the COPD effect, which is estimated to increase the log odds of hospitalization by about 0.35. The “informed” prior restricts the population effects to be positive but small, which tends to shrink the ITT_c effect toward zero – the overall ITT effect is identified, so positive effects in the other principal strata lead necessarily to smaller ITT_c estimates than unrestricted models.

The substantive conclusion is largely unchanged: there is only modest evidence for the effectiveness of the vaccine, particularly if we assume the direct effect of the reminder is similar in always-takers and compliers. This is unsurprising given the sample size of $n = 2,861$, weakness of the encouragement (Fig. 7), and relative rarity of the outcome (about 8% of the study participants were hospitalized). However, unlike Hirano et al. [2000] we conclude that the ITT_c effect is actually fairly robust to plausible deviations from the exclusion restrictions.

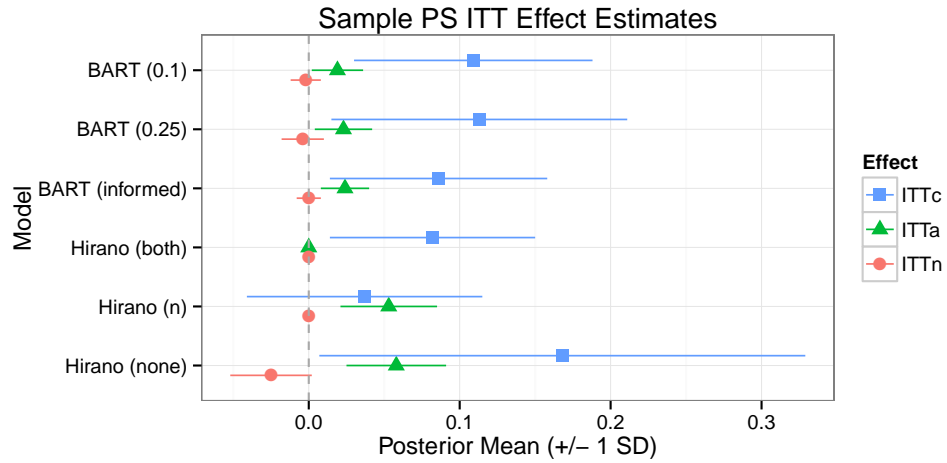


FIGURE 9. Posterior mean and standard deviations for the sample ITT effects in each principal stratum. “BART (0.1)” and “BART (0.25)” use the conditional prior centered at the exclusion restrictions with $\sigma_a = \sigma_n = 0.1$ or 0.25, respectively. “BART (informed)” includes all prior information. The “Hirano” models are from Hirano et al. [2000]; the “n” model assumes the exclusion restriction in never-takers, the “both” model imposes both exclusion restrictions and the “none” model has no exclusion restrictions.

APPENDIX C. IMPUTING PRINCIPAL STRATA MEMBERSHIP, POTENTIAL OUTCOMES, AND
 ITT EFFECTS

The full conditional distribution of principal strata membership is given by

(21)

$$\Pr(S_i = a \mid Y_i = y, T_i = t, Z_i = z, X_i = x, \phi, \theta) \propto \pi_A(x) \gamma_a^{11}(x)^y (1 - \gamma_a^{11}(x))^{1-y} \mathbb{1}(Z_i = T_i = 1)$$

(22)

$$\Pr(S_i = n \mid Y_i = y, T_i = t, Z_i = z, X_i = x, \phi, \theta) \propto \pi_N(x) \gamma_n^{00}(x)^y (1 - \gamma_n^{00}(x))^{1-y} \mathbb{1}(Z_i = T_i = 0)$$

(23)

$$\Pr(S_i = c \mid Y_i = y, T_i = t, Z_i = z, X_i = x, \phi, \theta) \propto \pi_C(x) \gamma_c^{tz}(x)^y (1 - \gamma_c^{tz}(x))^{1-y} \mathbb{1}(Z_i = T_i)$$

Note that observations with $Z_i = 1, T_i = 0$ ($Z_i = 0, T_i = 1$) are never takers (always takers) with probability one, and observations with $T_i = Z_i = 1$ ($T_i = Z_i = 0$) are either compliers or always takers (never takers).

Finally, given an observation's PS membership, the missing potential outcome is sampled from its full conditional, which is just

$$\Pr(Y_i(t, 1 - z) = 1 \mid T_i = t, Z_i = z, X_i = x, S_i = s, \phi, \theta) = \gamma_s^{t(1-z)}(x).$$

The ITT effects can be computed from the completed data using the following alternative expressions:

$$(24) \quad ITT_c = \frac{\sum_{i=1}^n \mathbb{1}(S_i = c) [Z_i \{Y_i - Y_i(0, 0)\} + (1 - Z_i) \{Y_i(1, 1) - Y_i\}]}{\sum_{i=1}^n \mathbb{1}(S_i = c)}$$

$$(25) \quad ITT_n = \frac{\sum_{i=1}^n \mathbb{1}(S_i = n) [Z_i \{Y_i - Y_i(0, 0)\} + (1 - Z_i) \{Y_i(0, 1) - Y_i\}]}{\sum_{i=1}^n \mathbb{1}(S_i = n)}$$

$$(26) \quad ITT_c = \frac{\sum_{i=1}^n \mathbb{1}(S_i = a) [Z_i \{Y_i - Y_i(1, 0)\} + (1 - Z_i) \{Y_i(1, 1) - Y_i\}]}{\sum_{i=1}^n \mathbb{1}(S_i = a)}$$

REFERENCES

- Securities and Exchange Commission, Accounting and Auditing Enforcement Releases. <http://www.sec.gov/divisions/enforce/friactions.shtml>, 2014.
- C.H Achen and W. Phillips Shively. *Cross-Level Inference*. University Of Chicago Press, first edition, 5 1995.
- J. Aldrich. How likelihood and identification went Bayesian. *International Statistical Review*, 70(1):79–98, 2002.
- J.D. Angrist, G.W. Imbens, and D.B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.
- J. Barnard, C.E. Frangakis, J.L. Hill, and D.B. Rubin. Principal stratification approach to broken randomized experiments: A case study of school choice vouchers in New York City. *Journal of the American Statistical Association*, 98(462):299–323, 2003.
- J.C.C. Chan and J.L. Tobias. Priors and posterior computation in linear endogenous variable models with imperfect instruments. *Journal of Applied Econometrics*, 2014. ISSN 1099-1255.
- H.A. Chipman, E.I. George, and R.E. McCulloch. Bayesian CART model search. *Journal of the American Statistical Association*, 93(443):935–948, 1998.
- H.A. Chipman, E.I. George, and R.E. McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.
- D.R. Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 215–242, 1958.
- M.J. Daniels and J.W. Hogan. *Missing data in longitudinal studies: Strategies for Bayesian modeling and sensitivity analysis*. CRC Press, 2008.
- A.P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B*, 41(1):1–31, 1979.
- P.M. Dechow, W. Ge, C.R. Larson, and R.G. Sloan. Predicting material accounting misstatements. *Contemporary accounting research*, 28(1):17–82, 2011.
- O.D. Duncan and B. Davis. An alternative to ecological correlation. *American Sociological Review*, 18(6):665–666, 1953.
- A. Dyck, N. Volchkova, and L. Zingales. The corporate governance role of the media: Evidence from Russia. *The Journal of Finance*, 63(3):1093–1135, 2008.
- A. Dyck, A. Morse, and L. Zingales. How pervasive is corporate fraud? Technical Report 2222608, Rotman School of Management, 2013.
- J.-P. Florens and A. Simoni. Bayesian identification and partial identification. *Unpublished manuscript. Toulouse School of Economics, Toulouse, France*, 2011.
- M. Fréchet. Sur les tableaux de corrélation dont les marges sont données. *Ann. Univ. Lyon Sect. A*, 9:53–77, 1951.
- R. Frisch. *Statistical Confluence Analysis by Means of Complete Regression Systems*. Universitetets Økonomiske Institut, 1934.
- A.E. Gelfand and S.K. Sahu. Identifiability, improper priors and Gibbs sampling for generalized linear models. *Journal of the American Statistical Association*, 94(445):247–253, 1999.
- S. Greenland. Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(2):267–306, 2005.

- P. Gustafson. Bayesian inference for partially identified models. *The International Journal of Biostatistics*, 6(2), 2010.
- P. Gustafson, A.E. Gelfand, S.K. Sahu, W.O. Johnson, T.E. Hanson, L. Joseph, and J. Lee. On model expansion, model contraction, identifiability and prior information: Two illustrative scenarios involving mismeasured variables [with comments and rejoinder]. *Statistical Science*, pages 111–140, 2005.
- T. Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica*, 11(1):pp. 1–12, 1943.
- T. Haavelmo. The probability approach in econometrics. *Supplement to Econometrica*, 12: iii–115, July 1944.
- J.J. Heckman. Simultaneous equation models with continuous and discrete endogenous variables and structural shifts. In S.M. Goldfeld and R.E. Quandt, editors, *Studies in nonlinear estimation*, pages 235–272. Balinger, 1976.
- J.J. Heckman. Dummy endogenous variables in a simultaneous equation system. *Econometrica*, 46:931–959, 1978.
- J.J. Heckman. Sample bias as a specification error. *Econometrica*, 47:153–161, 1979.
- J.L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):212–240, 2012.
- K. Hirano, G.W. Imbens, D.B. Rubin, and X.H. Zhou. Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, 1(1):69–88, March 2000.
- G. W. Imbens and D. B. Rubin. Estimating Outcome Distributions for Compliers in Instrumental Variables Models. *The Review of Economic Studies*, 64(4):555–574, October 1997a. ISSN 0034-6527. doi: 10.2307/2971731. URL <http://restud.oxfordjournals.org/lookup/doi/10.2307/2971731>.
- G.W. Imbens and D.B. Rubin. Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, pages 305–327, 1997b.
- S. Jayaraman and T. Milbourn. Whistle Blowing and CEO Compensation: The Qui Tam Statute. In *AFA 2011 Denver Meetings Paper*, 2010.
- J. Kadane. The role of identification in Bayesian theory. In S.E. Fienberg and A. Zellner, editors, *Studies in Bayesian econometrics and statistics*, chapter 5.2, pages 175–191. North-Holland, 1975.
- G. King. *A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data*. Princeton University Press, 2013.
- B. Kline and E. Tamer. Default Bayesian inference in a class of partially identified models. *manuscript, Northwestern University*, 2013.
- T.C. Koopmans. Identification problems in economic model construction. *Econometrica*, pages 125–144, 1949.
- T.C. Koopmans and O. Reiersol. The identification of structural characteristics. *Annals of Mathematical Statistics*, 21(2):165–181, 1950.
- T. Lancaster and G. Imbens. Case-control studies with contaminated controls. *Journal of Econometrics*, 71:145–160, 1996.
- R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data*. Wiley-Interscience, 2 edition, 9 2002. ISBN 9780471183860.
- C.F. Manski. *Identification problems in the social sciences*. Harvard University Press, 1995.
- C.F. Manski. *Partial identification of probability distributions*. Springer, 2003.

- C.F. Manski. *Identification for prediction and decision*. Harvard University Press, 2007.
- L.C. McCandless, P. Gustafson, and A. Levy. Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Statistics in medicine*, 26(11):2331–2347, 2007.
- L.C. McCandless, P. Gustafson, A.R. Levy, and S. Richardson. Hierarchical priors for bias parameters in Bayesian sensitivity analysis for unmeasured confounding. *Statistics in medicine*, 31(4):383–396, 2012.
- C.J. McDonald, S.L. Hui, and W.M. Tierney. Effects of computer reminders for influenza vaccination on morbidity during influenza epidemics. *MD computing: computers in medical practice*, 9(5):304–312, 1991.
- Gregory S Miller. The press as a watchdog for accounting fraud. *Journal of Accounting Research*, 44(5):1001–1033, 2006.
- N.-T. Molitor, N. Best, C. Jackson, and S. Richardson. Using Bayesian graphical models to model biases in observational studies and to combine multiple sources of data: application to low birth weight and water disinfection by-products. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(3):615–637, 2009.
- H.R. Moon and F. Schorfheide. Bayesian and frequentist inference in partially identified models. *Econometrica*, 80(2):755–782, 2012.
- Ari Pakman and Liam Paninski. Exact hamiltonian monte carlo for truncated multivariate gaussians. *Journal of Computational and Graphical Statistics*, 23(2):518–542, 2014.
- W. Phillips Shively. “Ecological” inference: The use of aggregate data to study individuals. *The American Political Science Review*, pages 1183–1196, 1969.
- D.J. Poirier. Partial observability in bivariate probit models. *Journal of Econometrics*, 12(2):209–217, 1980.
- R.L. Prentice and R. Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411, 1979.
- T.S. Richardson and J.M. Robins. Analysis of the binary instrumental variable model. In Rina Dechter, Hector Geffner, and Joseph Y. Halpern, editors, *Heuristics, Probability and Causality*, chapter 25. College Publications, 2010.
- T.S. Richardson, R.J. Evans, and J.M. Robins. Transparent parameterizations of models for potential outcomes. In J.M. Bernardo, M.J. Bayarri, James O. Berger, A. P. Dawid, David Heckerman, A.F.M Smith, and Mike West, editors, *Bayesian Statistics*, volume 9, pages 569–610. Oxford University Press, 2011.
- D.B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.
- E. San Martín and J. González. Bayesian identifiability: contributions to an inconclusive debate. *Chilean Journal of Statistics*, 1(2):69–91, 2010.
- E. Tamer. Partial identification in econometrics. *Annual Review of Economics*, 2:167–95, 2010.
- T.Y. Wang. Corporate securities fraud: insights from a new empirical framework. *Journal of Law, Economics and Organization*, 29(3):535–568, 2013.
- J.L. Zhang and D.B. Rubin. Estimation of causal effects via principal stratification when some outcomes are truncated by “death”. *Journal of Educational and Behavioral Statistics*, 28(4):353–368, 2003.